Gigabit Networking

Today's advanced networks based on HIPPI and fiber optics will soon be routine, and they will span continents.

●●●●●●●●●●●

Don E. Tolmie

Computer networks must become faster because the equipment that is being interconnected has increased in power and performance. Ethernet, with a 10-Mb/s speed, seemed awesome a few years ago, but is beginning to show its age as more machines are tied together and workstations attain the power of yesterday's mainframes.

Networks using gigabit speeds are just starting to become available and offer a whole new set of problems and potentials. The advanced networks proposed for supercomputers today will be the run-of-the-mill networks interconnecting workstations and other ADP equipment in the near future. Not only are the bandwidths increasing, but the distances spanned are going from machine-room-size to cross-country. Fiber optics is an enabling technology in this evolution, providing longer distances, higher data rates, and improved error rates.

Factors Driving Gigabit Networking

When networks were mainly used to carry key strokes between dumb terminals and mainframes, 9600 b/s was quite adequate; it was considerably faster than people could read. Today, it is more common to pass files and pictures between workstations, mainframes, and storage systems. The emphasis is on improving the user's productivity and avoiding network bottlenecks.

If a picture is worth a thousand words, then it probably also takes a thousand times the bandwidth to transfer that picture. People are not content with just pictures; presenting the computer output data in movie format (called visualization) is the newest craze, and offers even greater increases in user productivity. The potential bandwidth of the human eye-brain system has been calculated to be on the order of a few gigabits per second; hence, gigabit speeds should satisfy the individual user's needs for a while [1]. The networking factors of importance for visualization are raw speed and noninterference between data streams—if a visualization data stream is interrupted by another packet, the user sees a glitch, which is very distracting. Visualization sessions also tend to last for many seconds, compared to a single packet transfer, which may take only a few microseconds. Error control is also unique for visualization, since data in error is usually discarded rather than retransmitted.

As computers become faster, they also increase their appetite for data. A computer that is idle because of bottlenecks for input or output data is wasting computing cycles. A major factor is the bandwidth between the computer and its mass storage system. Mass storage systems used to be limited to single disks intimately attached to individual computer systems; today, the trend is for groups of disks to be shared among a group of networked workstations. The networking factors of importance for file transfers are raw speed and fairly large files; latency and interfering data streams are not major concerns.

An interesting idea gaining acceptance is the close coupling of many workstations to achieve the computing power of a supercomputer. Single central processing unit (CPU) supercomputers are running out of potential performance gains, due to the laws of physics limiting the speed of light and electrons. Performance gains in the future will be achieved by interconnecting many smaller computers and spreading the problem across all of them. This has humorously been termed "the attack of the killer micros." The networking factors of importance for remote procedure calls (RPCs) are raw speed, low cost (it should not cost more than the workstation), and low latency. After an initial large setup file, the information transferred tends to be mainly short data, control, and synchronizing messages.

High-speed interconnections used to be confined to machine rooms, with distances on the order of a few hundred feet. This was adequate when only machines needed to intercommunicate, but now the demands are for people to interact with the high-speed data directly through visualization. Placing the people close to the supercomputers is one solution, but has proven ineffective. We have found that a person will put up with an average

DON E. TOLMIE is Chairman of ANSI Task Group X3T9.3, responsible for HIPPI, IPI, and FC.

The Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36. This work was performed under the auspices of the U.S. Department of Energy.

system in their office rather than go 100 feet down the hall to a super system. It is not the walk that seems to bother people, it is the lack of their "environment." In their offices they have reference papers, the phone for important incoming calls, and other items necessary for the performance of their work. Hence, we must find ways to provide the necessary bandwidth in the person's office.

A major strength of supercomputers is that they can handle very large problems requiring massive amounts of high-speed memory. Problems like modeling the weather require vast amounts of memory and lots of high-speed number-crunching. A set of problems, called the "grand challenge" problems (e.g., weather forecasting), require the biggest and fastest machines available. Supercomputers are expensive to purchase and operate. It is not practical to place a supercomputer on each scientist's desk, nor to move all the scientists to a central computer site. The direction being taken is to provide high-speed access for the scientists, so that it seems as though they are sitting next to the machine. When the scientist is in Texas and the computer in California, you have a long-distance communications problem—or opportunity—depending on your point of view. The National Education and Research Network (NREN) has as its goal a 3-Gb/s backbone network across the nation within 10 years: this is ambitious. Part of the rationale is that computing is a national resource, which we must fully tap to stay competitive in today's global marketplace.

Data Communications and Telecommunications Cultures

One challenge is to marry the local area networks (LANs) used to interconnect the supercomputers with the long distance networks used for telecommunications. Data communications and telecommunications come from different backgrounds and cultures, with different goals, tools, and problems. This is not to say that one culture is right and the other is wrong; it is just that they are different. There must be a learning period for each culture to understand how the other works, and why it works the way it does.

One challenge is to marry the LANs used to interconnect the supercomputers with the long distance networks used for telecommunications.

Telecommunications and computer networks have traditionally used different techniques. Telecommunications networks have effectively used circuit switching and time-division multiplexing of many slow channels to a single fast channel; computer networks have used packet switching with datagrams, where each packet takes the total bandwidth of the media. Telecommunications networks have been very concerned with guaranteed bandwidth so that the latency is consistent, for example, not causing uneven time delays in speech traffic; computer networks were less worried about inconsistent delay, and more concerned with allowing use of all of the available bandwidth.

Now we are seeing the two "cultures" starting to merge. Computer networks need some of the guaranteed bandwidth of circuit-switching techniques to transmit isochronous video and voice among the end nodes. Likewise, telecommunications networks are becoming digital and using small packets, e.g., 53-octet (byte) cells in Asynchronous Transfer Mode (ATM) with the Synchronous Optical NETwork (SONET) for carrying multiple traffic streams. Telecommunications networks still need a call setup to load the address translation lookup tables in the route and possibly allocate bandwidth.

The messages in computer networks are normally sent in datagram mode. By this, we mean that a host sends a message to a destination when the media is available. The destination may reply, but the reply is a separate datagram message. The timing of the messages is indeterminate, as long as it is within some bounds. Bursty use of the media is the result. The protocols nominally run timers so that if a reply is not received within some defined time-out period, the original message is resent. This compensates for messages that are lost or garbled in the network. In contrast, the traffic for telecommunications networks has been mostly synchronous in nature: for example, voice or video. If the data does not arrive within stringent time constraints, it is useless and may as well be discarded.

Other potential problems associated with ATM include the fact that the cells do not include any error detection, e.g., parity, on the data portion of the cell. Cells may also be discarded by intermediate switches during overload conditions. Error recovery will probably not be done at the cell level, but at some group of associated cells, e.g., a packet. Hence, one bad or lost cell can cause the retransmission of a packet with a large number of cells.

In computer networks, i.e., data communications, a transmission normally takes the full bandwidth of the media for a short length of time. For example, a packet on Ethernet may be up to 1500 bytes long and is transmitted at the full 10-Mb/s media speed until all 1500 bytes have been transmitted. No other information flows during this time. A media access protocol is used to regulate when another host can transmit, ensuring that the messages do not interfere. In contrast, telecommunications networks multiplex many low-speed data paths onto a single high-speed media. As a rule, no one user gets the full media bandwidth.

Computer networks usually connect to hosts with both in and out paths, but nominally use only one direction at a time, i.e., either send or receive, but not both simultaneously. For example, in a bus or ring architecture like Ethernet or Fiber Distributed Data Interface (FDDI), which allows only a single message on the media, simultaneous transmission and reception of different messages is impossible. In contrast, telecommunications operations are usually full-duplex, with bandwidth available simultaneously in both directions, even though with voice traffic only one side is nominally used at a time.

Computer networks nominally use variable-size packets with defined maximum and minimum sizes. Large packet sizes are useful, as they incur the least amount of overhead for the amount of data transferred. Small packet sizes may be more efficient in that they take less time to transmit and incur less latency. The idea is to match the packet size to the amount of information to be sent. In contrast, most telecommunications systems operate on fixed-size units, for example, either allocated bandwidth or fixed cell sizes.

Conventional wisdom says that the less you "touch" the information to be transmitted, the lower the overhead, i.e., large packets are more effi-

• • • • •

That the standards process usually brings together the best and brightest people of many companies to work collectively on a problem.

cient. A LAN interface or bridge that can touch, or operate on, 20,000 packets/s is considered very fast, and effectively takes 50 μs for each packet. At 800 Mb/s, 50 μs translates into a 5-kbyte packet; anything smaller would result in an effective lower bandwidth, due to the 50 μs processing overhead. At the 2.4-Gb/s speed of SONET, a 53-octet cell takes less than 200 ns; hence, assembling and working with 53-octet cells is going to be a challenge at the higher SONET rates, e.g., approximately 5,600,000 cells/s at 2.4 Gb/s (OC-48).

As a computer network reaches saturation, a host sees longer delays in sending messages, but is usually not totally locked out. In essence, everyone on the network sees reduced performance, but no one is denied service. In contrast, when a telecommunications network saturates, it denies service to new users as they attempt to sign on. This protects the existing users but hinders new users. Telecommunications networks may throw away some cells during an overload, while this does not normally happen in computer networks.

When a local computer network is installed, the cost is mainly the capital cost of the equipment. Once purchased, it does not cost any more to fully load the network than it does to transmit one short message a day. In contrast, when using a telecommunications network, you are renting bandwidth and not purchasing equipment. A major goal of computer network vendors is to deliver data with a high degree of reliability. Once sold, the computer network vendor has minor interest in whether the equipment is used to its full capability or not. In contrast, the telecommunications industry is inter ested in keeping the channels as full of data as possible, since this is where they get their revenue.

Changes in Computer Networks

Computer and telecommunications networks are changing to accommodate higher speeds, longer distances, and new technology. The protocols useD in computer networks today were designed based on yesterday's technology. An example is Transmission Control Protocol/Internet Protocol (TCP/IP). This popular protocol was designed for long distance communications over the telephone system when error rates were on the order of 10^{-4} and speeds were in the 50-kb/s range. Now we are pushing gigabit speeds, and the transmission systems are much more reliable. It has been almost a chicken-and-egg situation. Until there were reliable high-speed communications paths available and in use, there was little incentive to build protocols to take advantage of them. XTP is an example of a protocol that is being developed with higher speeds in mind.

Some of the protocol changes to support higher data rates are obvious. For example, in TCP/IP the checksum is placed in the packet header, while more modern protocols put the checksum in a trailer so that it can be calculated as the data is being transmitted instead of requiring a separate pass through the data. Another example is packet size—when error rates are high, large packets are impractical, because the probability of a packet without errors is small. The higher data rates and longer distances also require a larger window size (the number of allowable unacknowledged packets), hence allowing more packets to be in transit.

Standards

The computing and telecommunications industries have become aware that hardware and software standards are necessary for future growth. No single company can provide all of the solutions, and interoperation with other vendors requires agreed-upon interfaces. Users are also demanding conformance to standards, so that they can purchase from multiple vendors and minimize their training costs. Some years ago, some people thought that standards stifled creativity. It is our observation that standards allow a company to invest a larger amount in their own areas of special expertise, with a smaller investment required to interface to the other vendors that conform to a standard. Otherwise, the cost of separate interfaces to each individual vendor may well outweigh the cost of the main business.

We have also seen that the standards process usually brings together the best and brightest people of many companies to work collectively on a problem. Design by committee really does work; the output of a standards committee is usually considerably more thorough and of higher quality than if one person or one company had done the complete job. We cannot say enough good things about the companies and individuals that support the voluntary standards efforts. In the gigabit computer networking arena, the High-Performance Parallel Interface (HIPPI) and Fiber Channel (FC) are examples of interfaces currently in the standards process. SONET and ATM are examples of standardization of higher speeds in the telecommunications industry. Protocol and software standards have also benefited from committee input.

High-Performance Parallel Interface

The HIPPI effort was started by the Los Alamos National Laboratory in early 1987. Our motivation was to have the vendors in the supercomputer community agree on a physical interface standard [2] so that separate interface adapters would not be required to connect to each vendor's proprietary interface. When we first took our proposal for an 800-Mb/s interface to the American National Standards Institute (ANSI) X3T9.3 Task Group, we were labeled as the "lunatic fringe"—who in the world would need anything that fast? Needless to say, we are no longer the lunatic fringe; in fact, some people say we aimed too low.

HIPPI was the first hardware standard in the super-computing arena. You may have heard of HIPPI previously as HSC or HPPI. The name was changed to avoid infringing on existing Digital Equipment Corporation (DEC) and Hewlett-Packard trademarks. Some of the initial X3T9.3 goals for the HIPPI physical-level interface (HIPPI-PH) included:

- A fire hose for moving data at 800 or 1600 Mb/s
- To get it done quickly, since we had immedi needs
- To use current technology—no new silicon required
- To avoid options
- To keep it simple

We achieved these goals, and the first HIPPI interfaces were delivered in late 1988. Since then, many vendors have implemented HIPPI on their products or are in the process of doing so. Currently, HIPPI is the interface of choice in the supercomputing arena.

HIPPI provides a point-to-point simplex data path; that is, it transfers in one direction only. Two back-to-back HIPPIs provide full-duplex or dual-simplex operation. 800 Mb/s is supported on one cable, and 1600 Mb/s requires two cables. The cables use twisted-pairs copper wires, are limited to 25 m in length, and are about 1/2-in. in diameter. Standard ECL drivers and receivers are used.

The hierarchy within HIPPI is:

- Connection—Must exist before data can be transferred
- Packet—Groups multiple bursts together into a logical entity
- Burst—Up to 1 or 2 kbytes, basic flow control unit, words within a burst are transferred synchronously with a 25-MHz clock; a checksum follows each burst
- Words—32 bits on 800-Mb/s HIPPI, 64 bits on 1600- Mb/s HIPPI, plus an odd parity bit for each byte in each word

HIPPI also provides a flow control mechanism that allows full bandwidth over many kilometers, for use with fiber optic extenders or across other networks such as SONET. Flow control is done on 1-kbyte or 2-kbyte bursts, decreasing the physical-level overhead. Error detection is done in a modular fashion on individual bytes and bursts, supporting very large (megabyte) packets in a consistent fashion. Error recovery is the responsibility of higher-layer protocols.

Networking at the physical layer is supported by HIPPI addressing and "connection" constructs. A common HIPPI network architecture uses a crossbar-type circuit switch (for example, a Network Systems Corporation PS32 Hub). It works much like a user's view of a telephone connection; that is, the HIPPI source provides a destination address (phone number), and the destination signals whether it can accept the connection (answers the phone). Once a connection is made, one or more packets of data may be passed without further interaction with the switch, i.e., the overhead is only while the connection is being completed. Either end may hang up, terminating the connection.

The suite of HIPPI documents has expanded beyond the HIPPI-PH described above. The HIPPI switch control (HIPPI-SC) defines how physical-layer switches operate and are addressed. The HIPPI framing protocol (HIPPI-FP) operates much like a data link layer, breaking large packets up into smaller bursts for transfer across HIPPI-PH. HIPPI-FP also specifies a header, describing to whom the packet belongs and where the data is located in the packet. Multiple protocols are supported above HIPPI-FP. IEEE 802.2 link encapsulation (HIPPI-LE) provides a mapping to the IEEE 802.2 data link for support of common network protocols such as TCP/IP. The HIPPI memory interface (HIPPI-MI) provides commands for reading and writing memory systems attached through HIPPI. A mapping to the Intelligent Peripheral Interface (IPI-3) command sets for disks and tapes is also supported, and is currently being used for stripped disk products.

Serial-HIPPI

The X3T9.3 Task Group wanted to use fiber optics when the HIPPI project was started, but at that time felt that the technology was not mature enough for the speeds we needed. Since we had an immediate need, the HIPPI interface was specified with copper twisted-pair cables limited to 25-m distances. Longer distances were needed for many applications, and this resulted in an *ad hoc* project to develop a "HIPPI extension cord" called Serial-HIPPI [3].

The purpose of Serial-HIPPI is to extend the physical range of HIPPI beyond 25 m and replace the parallel HIPPI cable by a single metallic or fiber cable. Other goals were to provide a low-error-rate link to support distances up to 10 km using parts available from multiple vendors. The Serial-HIPPI was to be transparent to the end systems; the only difference they would see would be an additional latency due to the time of flight.

A major task in Serial-HIPPI was picking the coding scheme for the serial stream. The contenders were 4b/5b as used in FDDI, 8b/10b as used in FC, 8b/10b plus forward error correction code (FEC), scrambling, and 20b/24b. The final decision was between 8b/10b + FEC and 20b/24b.

The 8b/10b + FEC scheme took a 32-bit HIPPI word with the four parity bits, added the HIPPI control signals, and then encoded it with an 8b/10b code. An eight-bit FEC code was then generated for this encoded block, and the FEC bits, along with the complement of the FEC bits for DC balance, were interspersed within the block along with synchronizing bits. This scheme allows correction of any single bit error in the block and detection of all double bit errors. A penalty is that the FEC added to the bandwidth, so that a 1.5-Gbaud serial signal was needed. A benefit is that an inexpensive laser with a higher bit error rate (BER) could be used in a link and still achieve a low system BER.

The 20b/24b scheme uses a running count of the number of ones and zeros transmitted, plus knowledge of the number of ones in the next 20-bit code group, to decide whether or not the next 20-bit code group should be inverted. This results in a DC-balanced code, but there may be up to 33 contiguous bits without a transition. The 20b/24b scheme uses a 1.2-Gbaud serial signal. A benefit of the scheme is that the clock recovery phase lock loop operates with a consistent update period, making it very stable and easy to implement.

Choosing between the 8b/10b + FEC and 20b/24b schemes was difficult. Proponents of each were knowledgeable and presented their cases well. A question that we had difficulty answering was, "What will be the most predominat error mechanism in operational links: single-bit random errors or burst errors?" The decision was finally made by a narrow margin to accept the 20b/24b scheme. A factor that influenced this decision included the fact that the predominant errors the participants had seen in similar links were bursty in nature, often caused by power supply fluctuations induced by power line disturbances. They stated that their links either ran without errors or were the equivalent of totally dead.

It was felt that the single-bit random errors were a real case, but occurred mainly when a link was being operated at its maximum limit, i.e., close to the noise

• • • • •

The purpose of Serial-HIPPI is to extend the physical range of HIPPI beyond 25 m and replace the parallel HIPPI cable by a single floor. Running at the maximum allowable distance is very important in the telecommunications world, where the distances require repeaters, and the fewer the repeaters the cheaper the link. This is not normally the case for computer links. It was felt that most Serial-HIPPI links would be about 1 km or less, and only a few would even approach 10 km. Hence, a little extra power could be used to escape the noise floor and avoid the random errors.

• • • • •

It should be noted that the emphasis of the CASA testbed is not computer networking, but applications.

HIPPI already includes error control information in the form of odd-byte parity and an even-parity checksum over all the words in a burst. The 20b/24b scheme arranged the HIPPI bits in a fashion so that one, two, or three bit errors would be detected by the 20b/24b coding or HIPPI checks. The possibility of undetected errors is extremely small. Also, the baud rate for the 8b/10b + FEC was 1.5 Gbaud, 25% higher than for the 20b/24b. It was felt that the higher baud rate would cost more, and the higher speed would also make it more error-prone, defeating its intent.

A chipset is being developed by Hewlett-Packard to implement the Serial-HIPPI 20b/24b coding. This coding scheme had already been chosen by the Scalable Coherent Interface (SCI) for their serial links. Using common parts should result in larger volumes and lower chipset costs. Serial-HIPPI specifies a baud rate of 1.2 Gbaud and a distance of up to 10 km using 9-µm single-mode fiber. The optical transmitter is specified as a pigtailed laser with a center wavelength between 1285 and 1330 nm, and a mean launch power of -9 to -6 dBm. The optical input to the receiver is specified as -22 to -6 dBm. The optical connector is either the FC/PC, or Super FC/PC, with a mean connector loss of less than 0.25 dB. The minimum unallocated optical power margin is +2.6 dB.

Serial-HIPPI also specifies copper coaxial cables for short distances. It was felt that coax would provide low cost, as well as a small connector footprint when compared to the standard HIPPI-PH 100-pin connector. Coaxial cable would be useful for intra-cabinet and local distribution, e.g., interconnecting a cluster of workstations. A motherboard/daughterboard packaging would allow the final drive elements, optical or electrical, to be tailored to the situation. 50-ohm coaxial cable is specified for limited distances, with or without an equalization circuit. Distances of up to 9 m are supported with RG 174U coax, and up to 36 m with RG 8/U. These distances could probably be improved with a better equalizer circuit.

A BER of $\leq 10^{-12}$ is specified for Serial-HIPPI; actual BERs are expected to be considerably better. A problem with specifying a lower BER is the time required to test a unit for compliance—possibly too long to be practical.

1600-Mb/s HIPPI, which uses a parallel 64-bit data word instead of the 32 bits of the 800-Mb/s version, is also supported by using two 800-Mb/s Serial-HIPPI units and two fibers in each direction. A synchronizing pattern is available for time-aligning the two units.

NREN and CASA Testbed

Supercomputers have proven to be very effective for simulating physical phenomena. Congress, in an attempt to increase the effectiveness of U. S. scientists and engineers, is pushing the NREN, with a goal of a coast-to-coast 3000-Mb/s computer network backbone [4, 5]. If you cannot move the users to the computers, then make the computers available to the users as if they were adjacent. There is a lot of research and testing going on to make the NREN a reality within the time goal. Los Alamos is participating in the CASA testbed. HIPPI is also being used heavily in the testbeds.

The Los Alamos National Laboratory is part of the NREN CASA gigabit testbed. CASA also includes the San Diego Supercomputer Center, Caltech, and the Jet Propulsion Lab (JPL) in Los Angeles. HIPPI with crossbar switches will be used at each of these sites as the LAN interconnecting the high-performance systems. SONET will be used as the long-distance physical-layer transport media. TCP/IP will be used as the transmission protocol.

It should be noted that the emphasis of the CASA testbed is not computer networking, but applications. That is, can we successfully run problems spread across supercomputers separated by large distances, and also have the users separated from the machines by large distances? For example, there are plans to run a global climate model that simultaneously uses the Cray YMP at the San Diego Supercomputer Center and the Connection Machine at Los Alamos, with the user at any CASA site.

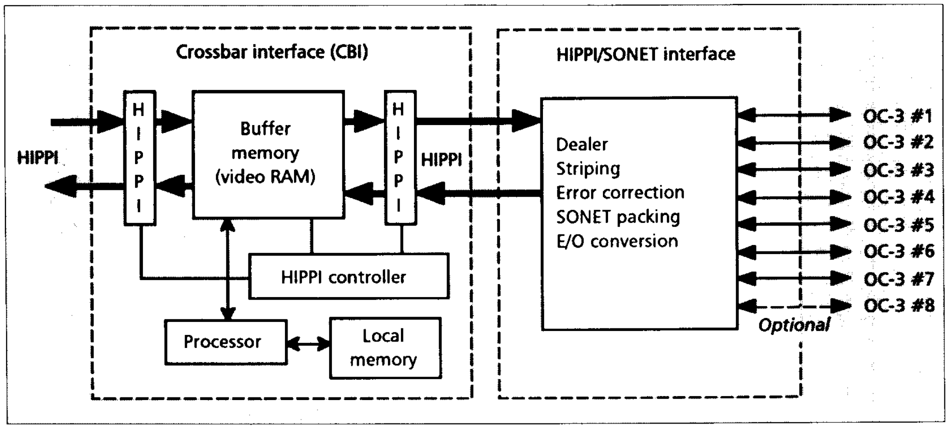
A question that we will be trying to answer is, how well can the problem be split across the widely spaced machines? The major difference that will be seen is the long communications delay introduced by the approximately 2000 km, or 10 ms, between the sites. Another research topic is the effective error rate: in practice, could the SONET BER really be as bad as 10^{-10}/hop, and if so how is this going to affect the end-to-end operation of the network? A BER of 10^{-10}/hop (with many hops) at 1 Gb/s translates into one or more errors per second, clearly not adequate for computer-to-computer operations.

Marrying Data Communications and Telecommunications

We are seeing a marriage of convenience between data communications and telecommunications. To span larger distances, computer users have looked to the telecommunications industry, which has the long lines in place. The decision to use SONET in the CASA testbed for the long hauls was based on several factors. We found that the telecommunications suppliers were less than enthusiastic about providing the equivalent of dark fiber across the country; there was not much that was very interesting for them in this approach. Using SONET was more in line with the telecommunications industry directions, and hence more interesting to them.

SONET defines a point-to-point capability, adequate for early experiments. The longer-term solution will require information switching in the network, and will probably be based on ATM. ATM uses a 53-octet cell with a 48-byte payload. The telecommunications industry is actively working on ATM switches operating at gigabit speeds.

Today's Internet uses store-and-forward switches, scattered around the world, interconnected by point-to-point links. The links are in the process



■ Figure 1. *Crossbar and HIPPI-SONET Interfaces.*

of being upgraded from T1 to T3. For the next generation, rather than providing the switching external to the telecommunications network, it seems advantageous to use shared lines and the telecommunications switching capability for switching. An advantage of this approach is that dedicating OC-48 (2.4 Gb/s) links between distributed data communications switching nodes will probably be expensive, while providing the bandwidth on demand, or after a short setup, may be cheaper. Also, if switching is already built within the telecommunications networks, then why duplicate the function with a separate system? A system that shares the data communications with the telecommunications should be more cost effective for everyone.

Some of the things that will affect this marriage of data communications and telecommunications will include the following:

- The latencies involved in setting up the telecommunications circuits
- Scheduling high-bandwidth connections for such things as visualization
- Providing an adequate BER
- The charging algorithms used

The possibility of using ATM-based telecommunications network equipment for local computer networks is an interesting concept being investigated at several locations. Supporting multimedia applications requires isochronous data for voice and video, and telecommunications networks provide this support naturally. ATM does not come free, though. The 48-byte ATM payload is very small compared to normal data communications network packets of 1 to 32 kbytes. Hardware support must be provided to assemble collections of ATM cells into larger entities.

Telecommunications networks are circuit-oriented, i.e., a virtual circuit is allocated between the source and destination and identifiers assigned. This virtual circuit setup is avoided in most of today's computer networks. Computer networks today are based on datagrams, which do not require a circuit to be set up before the datagram is transmitted. In contrast, a communications network requests a virtual circuit path first, allocating bandwidth and setting up the virtual circuit and virtual path identifiers before any data can be moved.

Providing a high-bandwidth physical connection between two machines does not guarantee that they can use the bandwidth effectively. The influences of the protocols above the physical layer have a major effect. Existing upper-layer protocols were designed to operate with yesterday's physical layers. Now, rather than error rates of 10^{-4}, error rates of 10^{-9} are expected, largely due to the improvements gained by using fiber optic components. The distances and transfer rates also affect the protocol. The delay between California and New York is 30 ms, allowing 3000 packets of 1 kbytes each to be in transit. Window sizes, flow control, and error recovery at the higher speeds need to be addressed.

Mapping HIPPI to SONET

Los Alamos is planning to build a gateway between HIPPI and SONET for use in the CASA testbed. The funding is not finalized on this project, so only a conceptual design is done so far. Figure 1 shows a block diagram of the proposed interface.

The crossbar interface (CBI) was originally designed for use as an intelligent interface between HIPPI hosts and HIPPI-based crossbar switches, and has HIPPI interfaces on each side. The CBI provides buffering, flow and routing control, and a proven platform for the HIPPI-SONET interface. The CBI provides 4 Mbytes of buffering in each direction, for a total of 8 Mbytes of storage. The HIPPI protocol requires about 1 kbyte of buffering/km of distance, or 2 Mbytes for the approximately 2000 km between Los Alamos and southern California. The HIPPI data and protocol are carried over SONET at speeds up to the 800-Mb/s HIPPI rate.

The connections from the HIPPI-SONET Interface to the SONET equipment are multiple OC-3 lines, each operating at about 155 Mb/s. HIPPI words are transmitted over each OC-3 line. For example, the first word is sent over OC-3 #1, the second word over OC-3 #2, etc. Each HIPPI word already contains error control information in the form of odd-byte parity. Optionally, OC-3 #8 implements a checksum over all of the OC-3 lines, allowing error data on an OC-3 line to be reconstructed. This is almost exactly the same as the Redundant Array of Inexpensive Disks (RAID-3) scheme being used for high-performance striped disks.

The HIPPI-SONET Interface circuitry will operate with any number of OC-3 lines operational at one time. This allows for failing OC-3 lines, or low-cost low-performance interface units with few OC-3 lines.

• • • • •

The possibility of using ATM-based telecommunications network equipment for local computer networks is an interesting concept being investigated at several locations.

• • • • •

While HIPPI is more of a communications interface, Fiber Channel was also intended to address the need for a faster input/output (I/O) channel to support peripherals such as disks and tapes.

Fiber Channel

When the standardization effort for HIPPI started in 1987, ANSI Task Group X3T9.3 wanted to use fiber optics for the increased distance and EMI/RFI benefits. Unfortunately, the fiber optic technology was not mature enough at that time, so HIPPI was based on copper cables to meet the time and simplicity goals. FC is a follow-on to HIPPI, building on many of the ideas introduced with HIPPI. FC is also being developed in ANSI Task Group X3T9.3.

While HIPPI is more of a communications interface, Fiber Channel was also intended to address the need for a faster input/output (I/O) channel to support peripherals such as disks and tapes. FC is structured to support the IPI-3 command sets for disk and tape, Small Computer System Interface (SCSI) command sets, IBM S/370 Block Multiplexer commands, and HIPPI-FP packets, and includes a mapping to the IP of TCP/IP.

FC, like HIPPI, is also a point-to-point interface, but is more general and supports more types of transfers. FC is more of an "all things to all people" type of interface. In the long run, FC will provide more capability than HIPPI, but its generality also produces more complexity, which in turn makes it harder to specify and implement. HIPPI can be built with standard off-the-shelf parts, while an effective FC implementation will require custom silicon. Also note that large-scale integration (LSI) chipsets are currently available for HIPPI designs.

Where options were avoided in HIPPI, FC is full of options. For example, FC supports four speeds, with data transfer rates of 100, 200, 400, and 800 Mb/s, corresponding to 132-, 266-, 531-, and 1062.5-Mbaud serial signaling rates. The FC media may be single-mode fiber, two sizes of multimode fiber, or even inexpensive copper coax cable for short distances. Optical transmitters may be light-emitting diodes (LEDs) or lasers. Combinations of the above are specified for different speeds and distances.

When HIPPI operates in a datagram mode, the higher-layer protocols worry about error recovery and retransmission. HIPPI also limits transfers to a single packet at a time, where the packet may be of any size. In contrast, FC supports three classes of service: Class 1— dedicated connection, guaranteed delivery, and frames received in transmitted order; Class 2—virtual connection, guaranteed delivery, frames may be reordered, frame-switching, and buffer-to-buffer flow control; and Class 3— datagrams, delivery, and frame ordering not guaranteed.

Class 1 is seen as very useful for visualization, where a dedicated connection may exist for long periods of time, and interference from other data streams is undesirable. Class 2 will probably be used heavily for traditional I/O transfers, where multiple transfers are open at one time with frames from the different transfers multiplexed on a single fiber. Class 3 can be used with traditional communications protocols where recovery and reordering are already handled in the upper-layer protocols, and connection setup times must be avoided.

FC is structured into four layers for ease of understanding and documentation. FC-0 specifies the physical layer, with the serial drivers, receivers, media, etc. FC-1 specifies the 8b/10b encoding and decoding scheme used to encode the data into a DC-balanced serial bit stream. FC-1 also defines special symbols for such things as Idle, SOF, EOF, etc. FC-2 defines the framing, e.g., where the address, control, data, and check fields are located and what they mean. FC-3 defines common services such as hunt groups, multicast, and stripping a single packet across multiple FC-0/1/2 combinations for higher bandwidth. FC-4s are the mappings to higher-layer protocols, e.g., to the IPI-3 command sets for disk and tape. The FC-0, FC-1, and FC-2 layers have been combined into the FC physical layer (FC-PH) document [6].

The logical hierarchy within FC is:

- Operation—Logical construct to identify and group things for an upper-layer protocol
- Exchange—Group of sequences, normally related to I/O control blocks
- Sequence—Unidirectional group of frames
- Frame—Basic transfer unit; contains header with addresses, control, offsets, etc., up to 2 kbytes of data, and a CRC checksum. The frame is the basic flow control unit; words within a frame are contiguous and transmitted in a synchronous fashion.

Identifier and offset fields are contained within each frame's header, allowing the receiving port to place the data in the proper place in memory, hopefully eliminating the need for data copies in the receiving computer. Considerable work has gone into providing multiple levels of indirection so that the individual frames can be handled by state machines implemented in silicon, rather than by a general-purpose processor. The feeling is that this is mandatory if we are to keep up with the data transfer rate, multiplexed frames, and the variety of applications.

For 10-km distances, FC-PH specifies 1300-nm lasers and single-mode fiber for all of the data rates. 50-μm multimode fiber and 780-nm lasers (CD ROM lasers) may be used at 531 Mbauds for distances up to 1 km, and for distances up to 2 km for the slower rates. 62.5-μm multimode fiber and 1330-nm LEDs may be used for distances up to 500 m with the 133- or 266-Mbaud rates.

FC-PH also specifies an electrical interface for limited distances using 75-ohm CATV or miniature coaxial cables. The cables are transformer-coupled, driven by ECL circuits, and use an equalizer at the receiver. The distances supported with RG 6/U-1 or RG 59/U-1 CATV coaxial cables are 25, 50, 75, and 100 m for the 133-, 266-, 531-, and 1063-Mbaud rates, respectively. Likewise, the distances using RG 179B/U miniature coaxial cables are 10, 20, 30, and 40 m. IEEE 802.5 shielded twisted-pair (STP) cable may also be used for distances of 50 and 100 m with the 133- and 266- Mbaud rates.

FDDI Follow-On LAN

Another interface addressing the gigabit-per-second LAN and wide area network (WAN) need is currently in the early development stages in ANSI Task Group X3T9.5. The FDDI Follow-On LAN (FFOL) is intended to support remote file access, image transfer, video, video conferencing, voice, multimedia applications, transactions processing, and low-latency real-time applications. FFOL is being designed to provide sufficient bandwidth to act as a backbone for multiple FDDI 100-Mb/s ring networks and allow efficient interconnection to

WANs such as Switched Multimegabit Data Service (SMDS) and Broadband Integrated Services Digital Network (B-ISDN) [7].

FFOL intends to use the existing FDDI cable plants with both multimode and single-mode fiber. Data rates will be scalable from STS-3 to at least STS-48 in the SONET/Synchronous Digital Hierarchy (SDH). Physical topologies of rings and trees will be supported, with maximum link distances of 10 km and a maximum network distance of 100 km.

Network Architectures

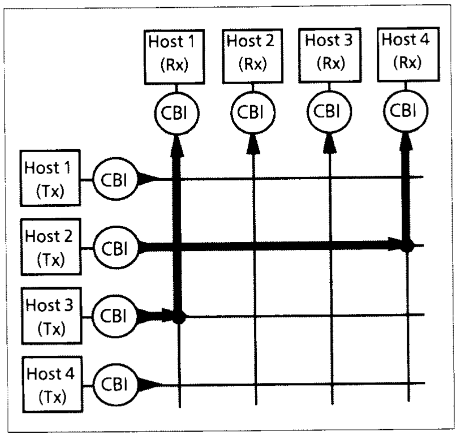
HIPPI and FC provide point-to-point connections that can be used as the basic building blocks for computer networks. Different types of network architectures are appropriate for different applications. HIPPI and FC lend themselves to circuit-switched, ring, and tree architectures.

HIPPI and FC provide point-to-point connections that can be used as the basic building blocks for computer networks.

Circuit-Switched Architectures

For comparison, circuit switching is the user's view of the telephone system today. That is, your call is separate and independent from someone else's call, even though you are both using the same circuited-switch hardware. The separate but independent nature of circuit switching is one of the requirements for visualization. The Los Alamos National Laboratory is prototyping a circuit-switching architecture called the Multiple Crossbar Network.

Figure 2 shows a 4 x 4 crossbar switch interconnecting four hosts. Note that connections exist for simultaneous transfers from Host 2 to Host 4, and from Host 3 to Host 1. The "CBI" nodes are the CBIs discussed earlier and shown in detail in Fig. 1. The CBIs would perform such functions as data buffering, flow control, switch access generation, address resolution, security checking, and low-level protocols. The CBIs are very similar to the CABs for the Carnegie-Mellon NECTAR project being developed by Network Systems Corporation. The CBI is also the building block for the HIPPI-SONET Interface described earlier.



■ Figure 2. *Circuit-switched architecture.*

The circuit-switched components run at the basic channel rate, and obtain a high total bandwidth by allowing multiple channels to be active simultaneously. For example, an 8 x 8 circuit switch for HIPPI would have each channel running at 800 Mb/s, the circuits within the switch running at 800 Mb/s, and a total bandwidth of 6.4 Gb/s. In use, one mainframe may be sending data to a visualization station, while another mainframe is reading data from a disk system, with both simultaneously transferring data at 800-Mb/s rates.

Normally, once a connection is completed, the channel operates as if there were no switch involved. That is, delays may occur on circuit setup, but no delays, other than circuit delays, are encountered once the connection is completed.

Circuit switches use different access control mechanisms than traditional bus or ring computer network architectures. Namely, if a source on a switch finds that its requested destination is busy, and the source has data for a different destination, the source can try sending to the second destination. With a bus or ring, if the media was busy, you could not send even if you had data for another destination.

Camp-on features may also be used to hang a source waiting for a specific destination to complete. Call-queuing schemes have also been proposed for connection setups. Switch systems need to watch out for hung channels and channel hogs. In the absence of a busy destination, setting up a circuit may take from less than 1 µs to 1 ms, depending on the switch size and connection control circuitry. Once completed, delays through the switch from a few nanoseconds to 1 µs may be encountered.

While a ring or bus system may grow indefinitely, one attachment at a time, circuit switches grow in major increments. For example, if you are using an 8 x 8 switch and want to add a ninth element, then you have to buy a whole other 8 x 8 switch and interconnect the switches. Switch architectures are often square, e.g., crossbars, but may be tailored to a variety of applications. For example, a local switch may interconnect several workstations but have only one connection to the main switch, supporting only one mainframe-to-workstation transfer at a time.

There are advantages to large switches, e.g., up to 4096 connections, and to small modular switches, e.g., 8 x 8 or 32 x 32, and vendors are building both. Some of the early uses may give us some guidelines on the best way to apply switches.

Ring Architectures

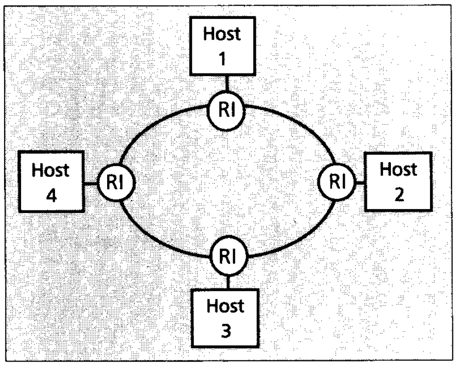
Computer networks based on ring architectures provide a single data path, which is shared by all of the attachments. This single data path limits the total bandwidth, but does give a natural broadcast capability. Bus access is usually determined by token passing or time slots. An advantage of rings is that it is usually fairly easy to add one more station. FDDI is an example of a ring network running at 100 Mb/s [8]. There are also some proposals for rings that allow multiple messages to simultaneously reside on the ring, increasing the potential bandwidth [9].

Figure 3 shows a ring network interconnecting four hosts. The ring interfaces (RIs) perform such functions as data buffering, ring access, security checking, and low-level protocols.

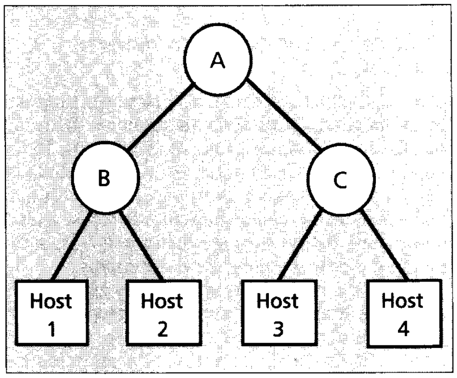
Fiber Channel-based rings are being considered for connecting peripherals, e.g., disks, to mainframes. In this environment, the limitation of a single data path is not critical, since the mainframe is normally the single generator and user of the data. It is envisioned that these rings would be cheaper than a circuit-switched architecture.

● ● ● ● ●

Fiber optics is an enabling technology, providing the longer distances, higher bandwidths, and improved error performance necessary for the next generation of network equipment.



■ Figure 3. *Ring architecture.*



■ Figure 4. *Tree architecture.*

Tree Architectures

Tree architectures, as shown in Fig. 4, allow simultaneous transfers as long as there is no contention, i.e., two messages trying to use a common link simultaneously. For example, Host 1 can transmit to Host 2 while, simultaneously, Host 3 is transmitting to Host 4. But if Host 1 is transmitting to Host 4, then Hosts 2 and 3 are locked out. Advantages of tree architectures are the simple and regular structure of the switching nodes (A, B, and C in Fig. 4) and the modular way they can be interconnected. Tree architectures are not very common in today's computer networks, but are being proposed for low-cost FC systems.

Conclusions

Computer networks operating at gigabit-per-second transfer rates are seen as necessary for many applications, and gigabit networks are becoming available. HIPPI, FC, and FFOL will provide some of the basic building blocks for these networks. Gigabit-speed LANs will be interconnected via WANs and telecommunications networks using SONET and ATM. These combined networks will provide increased user productivity and long distance data sharing. Fiber optics is an enabling technology, providing the longer distances, higher bandwidths, and improved error performance necessary for the next generation of network equipment.

Acknowledgments

The author wishes to thank Bill Grant, Wally St. John, and Steve Tenbrink for reviewing this article and providing valuable input.

References

[1] P. R. Rupert, "What's Driving Gigabit/Sec Channels?" *High-Speed Networks and Channels, SPIE*, vol. 1577, Sept. 4-6, 1991.

[2] ANSI, "High-Performance Parallel Interfac—Mechanical, Electrical, and Signaling Protocol Specification (HIPPI-PH)," Amer. Nat'l. Std. X3.183-1991.

[3] "Serial-HIPPI Specification, Revision 1.0, Serial HIPPI Implementers Group," available via anonymous ftp as file pub/hippi/serial_hippi_1.0.ps on hplsci.hpl.hp.com (15.255.176.57), or as file hippi/serial_hippi_1.0.ps on nsco.network.com (129.191.1.1).

[4] "Gigabit Network Testbeds," *IEEE Comp.*, vol. 23, no. 9, pp. 77-80, Sept. 1990.

[5] "Science and Business (Gigabit Connection)," *Sci. Amer.*, pp. 118-120, Oct. 1990.

[6] "Fibre Channel—Physical Layer (FC-PH)," ANSI X3T9.3 Wkg. Doc., rev. 2.2, Jan. 24, 1992.

[7] R.L. Fink and F.E. Ross, "Following the Fiber Distributed Data Interface," *IEEE Network*, Mar. 1992.

[8] F.E. Ross, "FDDI—A Tutorial," *IEEE Commun. Mag.*, vol. 24, no. 5, pp. 10-17, May 1985.

[9] I. Didon and Y. Ofek, "MetaRing—A Full-Duplex Ring with Fairness and Spatial Reuse," *INFOCOM '90*, pp. 969-981, 1990.

Biography

DON E. TOLMIE received a B.S.E.E. degree from New Mexico State University in 1959 and an M.S.E.E. degree from University of California, Berkeley, in 1961. He joined the Los Alamos National Laboratory in 1959 as a Technical Staff Member, and has been involved with the networking of supercomputers for almost 20 years. His current task is defining the next-generation computer network to support higher speeds and visualization, and working with vendors to provide the appropriate products. He has been involved in computer interface standards activities for over 10 years, and is currently Chairman of ANSI Task Group X3T9.3, responsible for HIPPI, IPI, and FC. He is the initiator and leader of the HIPPI effort, the current interface of choice for networking supercomputers and associated equipment at 800-Mb/s speeds.